

Approximate Range Counting Revisited*

Saladi Rahul

Department of Computer Science and Engineering
University of Minnesota
sala0198@umn.edu

Abstract

This work presents several new results on approximate range counting. For a given query, if the actual count is k , then the data structures in this paper output a value, τ , lying in the range $[(1 - \varepsilon)k, (1 + \varepsilon)k]$. The main results are the following:

- A new technique for efficiently solving any approximate range counting problem is presented. This technique can be viewed as an enhancement of Aronov and Har-Peled's technique [*SIAM Journal of Computing*, 2008]. The key reasons are the following:
 - (1) The new technique is sensitive to the value of k : As an application, this work presents a structure for *approximate halfspace range counting* in $\mathbb{R}^d, d \geq 4$ which occupies $O(n)$ space and solves the query in $\tilde{O}((n/k)^{1-1/\lfloor d/2 \rfloor})^1$ time. When $k = \Theta(n)$, then the query time is $\tilde{O}(1)$. The answer is correct with high probability.
 - (2) The new technique handles colored range searching problems: As an application, the *orthogonal colored range counting* problem is solved. Existing structures for exact counting use $O(n^d)$ space to answer the query in $O(\text{polylog } n)$ query time. Improving these bounds substantially would require improving the best exponent of matrix multiplication. Therefore, if one is willing for an approximation, an attractive result is obtained: an $O(n \text{ polylog } n)$ space data structure and an $O(\text{polylog } n)$ query time algorithm.
- An optimal solution for some *approximate rectangle stabbing counting* problems in \mathbb{R}^2 . This is achieved by a non-trivial reduction to planar point location.
- Finally, an efficient solution is obtained for *3-sided orthogonal colored range counting*. The result is obtained by a non-trivial combination of two different types of random sampling techniques and a reduction to non-colored range searching problem.

1 Introduction

1.1 Standard geometric intersection query (Standard GIQ)

In a standard *geometric intersection query (GIQ)*, a set S of n geometric objects in \mathbb{R}^d is preprocessed into an efficient data structure so that for any geometric query object, q , all the objects in S intersected by q can be reported (*reporting query*) or counted (*counting query*) quickly. In an *approximate counting query*, an approximate value of the number of objects in S intersecting q has to be reported; specifically, any value τ which lies in the range $[(1 - \varepsilon)k, (1 + \varepsilon)k]$, where $k = |S \cap q|$ and $\varepsilon \in (0, 1)$. In an *emptiness query*, we want to decide if $|S \cap q| = 0$ or not. Notice that the

*This research was partly supported by a Doctoral Dissertation Fellowship (DDF) from the Graduate School of University of Minnesota.

¹The symbol \tilde{O} hides the dependency on the $O(\text{polylog } (n/k))$ term and the ε term.

approximate counting query is *at least* as hard as the emptiness query: When $k = 0$, we do not tolerate any error. Therefore, a natural goal while solving an approximate counting query is to *match* the space and the query time bounds of the corresponding emptiness query.

Approximate counting queries is the focus of this paper. Arguably, the three most popular types of *GIQ* problems are (i) *orthogonal range searching* (S contains points, q is an axes-parallel rectangle) [1, 2, 11, 19, 22, 41], (ii) *rectangle stabbing* (S contains axes-parallel rectangles, q is a point) [3, 8, 9, 10, 21, 29], and (iii) *halfspace range searching* (S contains points, q is a halfspace) [5, 7, 13, 17, 36, 35]. While approximate counting queries have been well studied for (i) and (iii), there has been no concrete study of (ii).

A Brief History of Approximate Range Counting: Most of the focus in the early days of research on approximate range counting was on halfspace range queries. Starting from the work of Aronov and Har-Peled [15], there was a series of results by Kaplan and Sharir [32], Afshani and Chan [4], Aronov, Har-Peled and Sharir [16], and Kaplan, Ramos and Sharir [30]. These papers dealt with either halfspace range queries in low dimensional space ($d \leq 3$) or high dimensional space ($d \geq 4$). Later, Afshani, Hamilton and Zeh [6] obtained an optimal solution for a general class of problems which included halfspace range query in \mathbb{R}^3 , dominance query in \mathbb{R}^3 and 3-sided orthogonal range query in \mathbb{R}^2 . Interestingly, their results hold in the pointer machine model, the I/O-model and the cache-oblivious model as well. Two-dimensional orthogonal range searching query was studied by Nekrich [38], and Chan and Wilkinson [20] in the word RAM model. The ultimate goal in all these problems is to match the space and the query time bounds of their corresponding emptiness query.

Our results for standard GIQ problems: In this paper we study the halfspace range searching problem and the rectangle stabbing problem.

Approximate Halfspace Range Counting in $\mathbb{R}^d, d \geq 4$: We present a structure for halfspace range counting which is *sensitive* to the value of k . The data structure occupies $O(n)$ space and solves the query in $\tilde{O}\left((n/k)^{1-1/\lfloor d/2 \rfloor}\right)$ time. The answer is correct with high probability. When $k = \Theta(n)$, then the query time is $\tilde{O}(1)$, which is an attractive property to have. See Corollary 1 in Subsection 3.3 for a formal statement. Previously, such sensitive data structures were known only in $d = 2, 3$ [6]. In $\mathbb{R}^d, d \geq 4$, existing structures occupy $\tilde{O}(n)$ space and solve the query in $\tilde{O}\left(n^{1-1/\lfloor d/2 \rfloor}\right)$ time.

Approximate Rectangle Stabbing Counting in \mathbb{R}^2 : This paper initiates the concrete study of approximate rectangle stabbing counting. This specific problem is studied in the word-RAM model of computation. Consider the following two settings:

- (1) S contains 2-sided rectangles of the form $[x_1, \infty) \times [y_1, \infty)$. It is easy to see that this is a 2D dominance query. There is a *gap* between the 2D dominance counting query and the 2D dominance emptiness query. For 2D dominance counting query, Patrascu [39] gave a lower bound of $\Omega\left(\frac{\log n}{\log \log n}\right)$ query time for any data structure which uses $O(n \text{ polylog } n)$ space. On the other hand, for 2D dominance emptiness query, there is a linear space structure with query time $O(\log \log n)$.
- (2) S contains 3-sided rectangles of the form $[x_1, \infty) \times [y_1, y_2]$. This problem also has a gap between the counting query and the emptiness query. The bounds mentioned above for the counting query and the emptiness query hold true for this setting as well.

Our first result is a solution for approximate 2D dominance counting query and approximate 3-sided rectangle stabbing counting query whose bounds *match* their corresponding emptiness query: An $O(n/\varepsilon)$ size data structure for answering the query in $O(\log \log(n/\varepsilon k))$ time. See Theorem 2 for a formal statement. Adapting existing techniques (for e.g., Afshani, Hamilton and Zeh [6]) leads

to a solution for these problems with $\Theta((\log \log n)^2)$ query time, and with costlier dependency on ε in the space and the query time.

We do not study the case where S contains 4-sided rectangles of the form $[x_1, x_2] \times [y_1, y_2]$; because, this problem *does not* have a gap between the counting query and the emptiness query. For the emptiness query and the counting query, Patrascu [40] and Patrascu [39], respectively, gave a lower bound of $\Omega\left(\frac{\log n}{\log \log n}\right)$ query time for any data structure which uses $O(n \text{ polylog } n)$ space. JaJa, Mortensen and Shi [27] gave a linear space structure with matching query time for both the problems.

1.2 Colored-GIQ

Several practical applications have motivated the study of a more general class of *GIQ* problems, known as *colored-GIQ* problems [12, 24, 28, 31, 33, 34, 38, 42]. In this setting, the set S of n geometric objects in \mathbb{R}^d come aggregated in disjoint groups. Each group is assigned a unique color. Given a geometric query object, q , we are interested in reporting (*colored reporting query*) or counting (*colored counting query*) the colors which have at least one object intersected by q . Note that a standard *GIQ* problem is a special case of its corresponding colored-*GIQ* problem (assign each object in the standard *GIQ* problem a unique color). The most popular and well studied colored-*GIQ* problem is the *orthogonal colored range searching* problem: S is a set of n points in \mathbb{R}^d and q is an axes-parallel rectangle in \mathbb{R}^d . A motivating example for this problem would be the following database query: “How many countries have employees aged between 30 and 40 while earning more than 80,000 per year”. Each employee can be represented as a point (*age, salary*) and the query is represented as an axes-parallel orthogonal rectangle (unbounded in one direction) $[30, 40] \times [80,000, \infty)$. Each employee is assigned a color based on his nationality.

A general technique for hard counting problems: Unfortunately, for most colored counting queries the known space and query time bounds are very expensive. For example, for orthogonal colored range searching problem in \mathbb{R}^d , existing structures use $O(n^d)$ space to achieve polylogarithmic query time. Any substantial improvement in these bounds would require improving the best exponent of matrix multiplication [31]. Instead of an exact count, if one is willing to settle for an approximate count, then this paper presents a result with attractive bounds: an $O(n \text{ polylog } n)$ space data structure and an $O(\text{polylog } n)$ query time algorithm. See Corollary 2 in Subsection 3.3 for a formal statement.

In an *approximate colored counting query*, an approximate value of the number of colors in S intersecting q has to be reported; specifically, any value τ which lies in the range $[(1 - \varepsilon)k, (1 + \varepsilon)k]$, where k is the number of colors which have at least one object intersected by q . In Section 3 we present a general technique to reduce any approximate colored counting problem to a “small” number of its corresponding colored reporting queries for which usually faster solutions exist.

Is linear space and $\log n$ query time possible? There are some instances of colored-*GIQ* problems which are not “hard”. For example, for orthogonal colored range searching, there are two settings for which exact counting can be done using $O(n \text{ polylog } n)$ space and in $O(\text{polylog } n)$ query time: (a) points lying in \mathbb{R}^1 and the query is an interval $[x_1, x_2]$, and (b) points lying in \mathbb{R}^2 and the query is a 3-sided rectangle of the form $[x_1, x_2] \times [y, \infty)$. So, a natural question is whether by allowing approximation a linear space data structure and an $O(\log n)$ query time solution can be obtained for these problems? In this paper we show that it is indeed possible. Specifically, we study the setting in (b) as it is more challenging. Please see Theorem 5 for a formal statement. We note that Nekrich [38] presented an approximate solution for the same problem but with an approximation

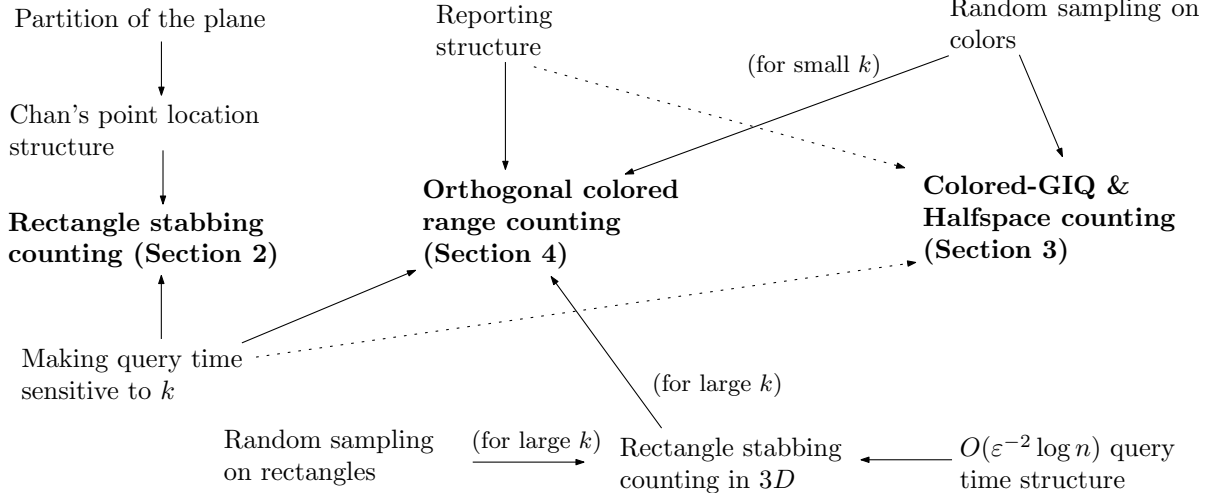


Figure 1: An overview of the techniques used (figure inspired from [20]).

factor of $(4 + \varepsilon)$, whereas we are interested in obtaining a tighter approximation factor of $(1 + \varepsilon)$.

1.3 Our techniques

This paper introduces some new ideas and also combines previous techniques in a non-trivial manner (see Figure 1 for an overview of the techniques used). The highlights are the following:

- A general technique for solving approximate counting of standard GIQ and colored-GIQ problems. Our technique can be viewed as an enhancement of Aronov and Har-Peled's approximate counting technique [15]. See Subsection 3.4 for details. We introduce the idea of performing random sampling on *colors* (instead of input objects) to approximately count the colors intersecting the query object.
- The result for approximate rectangle stabbing counting is obtained by a non-trivial reduction to planar point location.
- The result for approximate orthogonal colored range counting is obtained by a non-trivial combination of two different types of random sampling techniques and a reduction to non-colored range searching problem.

2 Approximate Rectangle Stabbing Counting in \mathbb{R}^2

In this section we study the approximate rectangle stabbing counting (*ARSC*) problem in \mathbb{R}^2 . The input set S is a set of n 3-sided rectangles in \mathbb{R}^2 and the query q is a point. We first prove the following result.

Theorem 1. *There exists a data structure of size $O\left(\frac{n}{\varepsilon}\right)$ which can solve approximate 3-sided rectangle stabbing counting (*ARSC*) problem in $O\left(\log \log \left(\frac{n}{\varepsilon}\right)\right)$ time.*

2.1 Outline of the solution

A brief outline of our solution. Given the set S of n 3-sided rectangles, the objective is to partition the entire plane into interior-disjoint rectangles R_{new} (rectangles can intersect only at the edges).

Each rectangle $r \in R_{new}$ will have a weight $w(r)$ associated with it. We impose the following two constraints on R_{new} :

1. $|R_{new}| = O(\varepsilon^{-1}n)$.
2. For any given query q , if $|S \cap q| = k$, then the rectangle $r \in R_{new}$ stabbed by the point $q(q_x, q_y)$ will have weight $w(r) \in [(1 - \varepsilon)k, (1 + \varepsilon)k]$.

We first study a simpler problem in Subsection 2.2 and then use the result to construct the set R_{new} in Subsection 2.3. Given the set R_{new} , the query algorithm is straightforward: In the preprocessing phase, based on R_{new} build the linear-size point location structure of Chan [18]. Given a query point q , locate the rectangle $r \in R_{new}$ containing q in $O(\log \log(\frac{n}{\varepsilon}))$ time and then report $w(r)$.

2.2 Partition of the real line

In this subsection we study a simpler problem of partitioning the real line to achieve certain desired properties.

Problem: We are given a set P of points lying on the real line (call it x -axis), which is initially empty. After that, a sequence of m update operations are performed on P . An update operation includes (a) inserting a new point into P , or (b) deleting an existing point from P . We also partition the real line into interior-disjoint intervals \mathcal{I} (intervals can touch only at the endpoints) and each interval $I \in \mathcal{I}$ will have a weight $w(I)$ associated with it. We impose the following two constraints on the set \mathcal{I} :

1. After each update to P , we are allowed to make *changes* to set \mathcal{I} . A change is either (a) inserting a new interval into \mathcal{I} , or (b) deleting an existing interval from \mathcal{I} . Let $ch(t)$ be the number of changes made to set \mathcal{I} after the t -th update to P , where $t \in [0, m]$. Then we want $\sum_{t=1}^m ch(t) = O(\varepsilon^{-1}m)$. In other words, the total number of changes to set \mathcal{I} is $O(\varepsilon^{-1}m)$.
2. After t updates to P , suppose we perform an approximate counting query on the set P . Given a query interval q of the form $(-\infty, q_x]$, let k be the number of points of P lying in q and let $I \in \mathcal{I}$ be the interval containing q_x . Then we want $w(I) \in [(1 - \varepsilon)k, (1 + \varepsilon)k]$.

Invariants: We will now show that it is indeed possible to build an algorithm which can satisfy both the constraints of set \mathcal{I} . We start with some definitions. After a sequence of t updates, let m_t be the number of points in P . Given a real number x , its *rank* is the number of points of P in $(-\infty, x]$. Note that by this definition, the rank of the i -th smallest point in P is i . Finally, define a variable ε' such that $\varepsilon = C\varepsilon'$, where C is a sufficiently large positive constant. The reason for the choice of parameter ε' will become clear from the analysis.

Notice that when $k < 1/\varepsilon$, then no approximation is allowed, i.e., an exact count has to be reported. To handle $k \geq 1/\varepsilon$, we try to maintain the “approximate” $\frac{1}{\varepsilon}(1 + \varepsilon)$ -th, $\frac{1}{\varepsilon}(1 + \varepsilon)^2$ -th, $\frac{1}{\varepsilon}(1 + \varepsilon)^3$ -th, \dots rank of P . The heart of the algorithm involves maintaining a list L whose entries maintain approximate ranks of P . After processing of each update, we force the list L to satisfy the following four invariants:

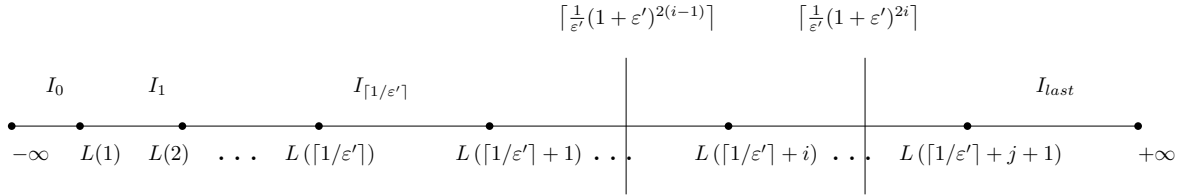
- Invariant 1 for $m_t \leq \lceil 1/\varepsilon' \rceil$: The number of entries in L will be m_t .
- Invariant 2 for $m_t \leq \lceil 1/\varepsilon' \rceil$: For $i \in [1, m_t]$, the i th entry, $L(i)$, stores the x -coordinate of the point of P whose rank is i . In other words, L is a *replica* of P .

- Invariant 3 for $m_t > \lceil 1/\varepsilon' \rceil$: The number of entries in L will be $\lceil 1/\varepsilon' \rceil + j + 1$, where j is the largest integer such that $m_t \geq \lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2j+1} \rceil$.
- Invariant 4 for $m_t > \lceil 1/\varepsilon' \rceil$: As in invariant 2, the first $\lceil 1/\varepsilon' \rceil$ entries of L are again a replica of smallest $\lceil 1/\varepsilon' \rceil$ ranks in P . The remaining $j + 1$ entries store approximate ranks. Specifically, for $i \in [1, j + 1]$, the rank of $L(\lceil 1/\varepsilon' \rceil + i)$ -th entry is in the range $(\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2(i-1)} \rceil, \lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i} \rceil)$.

Defining the partition: We now define the set of intervals $I_0, I_1, \dots, I_{last}$ in \mathcal{I} which will partition the real line. First, we define the *range* of each interval on the real line. $I_0 = (-\infty, L(1))$ and $I_{last} = [L(last), +\infty)$. For $i \in [1, last - 1]$, $I_i = [L(i), L(i + 1))$. Next, we define the *weight* assigned to each interval.

$$w(I_i) = \begin{cases} 0 & \text{if } i = 0 \\ i & \text{if } i \leq \lceil 1/\varepsilon' \rceil \\ \lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i-1} \rceil & \text{if } i > \lceil 1/\varepsilon' \rceil \end{cases}$$

The above definition misses one special case. When $m_t = 0$, then L is empty. Then \mathcal{I} will have only one interval $I_0 = (-\infty, +\infty)$ with $w(I_0) = 0$. In Section 5 of the appendix we prove that this definition of set \mathcal{I} satisfies the second constraint imposed on it.



Updating the list L : Before we describe the details of the update algorithm, we define a term *median position*. For $i \leq \lceil 1/\varepsilon' \rceil$, the median position of $L(i)$ will be the x -coordinate of the point of P with rank i . For $i \in [1, j + 1]$, the median position of $L(\lceil 1/\varepsilon' \rceil + i)$ will be the x -coordinate value of the point of P with rank $\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i-1} \rceil$. Notice that Invariant 4 allows each entry $L(\lceil 1/\varepsilon' \rceil + i)$ to deviate from its median position by a factor of roughly $(1 + \varepsilon')$.

Let the $(t + 1)$ -th update be *insertion* of a point into P , i.e., $m_{t+1} = m_t + 1$. The invariants are fixed as follows:

(1) *Updating existing entries in L :* Some of the entries in L might start violating Invariant 2 or Invariant 4. To fix this, each violating entry is set to its median position.

(2) *Creating a new last entry in L :* There are three scenarios under which a new entry will be included in L . First, when $m_t < \lceil 1/\varepsilon' \rceil$. Then $m_{t+1} \leq \lceil 1/\varepsilon' \rceil$. So we add a new entry $L(m_{t+1})$ whose value is set to its median position. This fixes the violation of Invariant 1 and 2. Second, when $m_t = \lceil 1/\varepsilon' \rceil$. Then $m_{t+1} = \lceil 1/\varepsilon' \rceil + 1 = \lceil \frac{1}{\varepsilon'}(1 + \varepsilon') \rceil$. Then we add a new entry $L(\lceil 1/\varepsilon' \rceil + 1)$ whose value is set to its median position. This fixes the violation of Invariant 3 and 4.

Finally, we look at the case where $m_t > \lceil 1/\varepsilon' \rceil$. Recall that j is the largest integer such that $m_t \geq \lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2j+1} \rceil$. After the insertion of a point into P , it might happen that j is *no longer* the largest integer such that $m_{t+1} \geq \lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2j+1} \rceil$. This implies that now $(j + 1)$ is the largest integer such that $m_{t+1} \geq \lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2(j+1)+1} \rceil$. Then we add a new entry $L(\lceil 1/\varepsilon' \rceil + j + 2)$ into the list L . The value of $L(\lceil 1/\varepsilon' \rceil + j + 2)$ is set to its median position. This fixes the violation of Invariant 3 and 4.

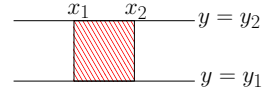
In Section 6 of the appendix, we show how to update the list L when a point is *deleted* from P . In Section 6, we also show that the total number of changes made to the list L (and hence to \mathcal{I}) is bounded by $O\left(\frac{m}{\varepsilon}\right)$. This proves that the set \mathcal{I} satisfies the first constraint as well.

2.3 Partition of the plane

Now we describe the construction of the set R_{new} . In the previous subsection, we built a data structure \mathcal{I} to maintain “a partition of the real line”. In this subsection, roughly speaking, we will make the data structure \mathcal{I} “persistent”. In the end, it will be clear that this persistent structure of \mathcal{I} is actually a partition of the plane and satisfies the two constraints on the set R_{new} .

The algorithm is a sweep line based approach on the set S of n 3-sided rectangles. Consider a horizontal line h which starts sweeping the plane upwards from $y = -\infty$. Initialize the structure built in the previous subsection to maintain a partition of the horizontal line h . Initially, the set P is empty and the set \mathcal{I} contains only one interval $(-\infty, +\infty)$ with weight 0. When the sweep line h visits the lower edge of a rectangle (say $r = [x_1, \infty) \times [y_1, y_2]$) in S , then a point with x -coordinate x_1 is inserted into P . On the other hand, when the sweep line h visits the upper edge of a rectangle in S , then the point corresponding to that rectangle is deleted from P . After each update to the set P , the set \mathcal{I} is also updated to ensure that it maintains the partition of the horizontal line h .

Construction of a rectangle: Every interval in \mathcal{I} *created* during the sweep corresponds to a rectangle in R_{new} . Consider an interval $[x_1, x_2] \in \mathcal{I}$. Assume y_1 be the y -coordinate of the sweep line when the interval was inserted into \mathcal{I} ; and let y_2 be the y -coordinate of the sweep line when the interval was deleted from \mathcal{I} . Then we create a rectangle $r_{new} = [x_1, x_2] \times [y_1, y_2]$ and add it to R_{new} . A weight $w(r_{new})$ is assigned to r_{new} which is the weight assigned to the interval $[x_1, x_2]$ when it was in \mathcal{I} .



One special case has to be taken care of. At the end of the sweep, there will be exactly one interval, $(-\infty, +\infty)$, in \mathcal{I} . This is the only interval in \mathcal{I} which gets inserted but does not get deleted. This interval is converted to a rectangle $(-\infty, +\infty) \times (y_{max}, +\infty)$ with weight 0 and added to R_{new} . Here y_{max} is the y -coordinate of the upper endpoint with the largest y -coordinate. With a moment's thought, it is clear that our construction ensures the two constraints imposed on the set R_{new} .

With some additional ideas it is actually possible to obtain the following stronger result. The proof of this theorem can be found in Section 7 of the appendix..

Theorem 2. *There exists a data structure of size $O\left(\frac{n}{\varepsilon}\right)$ which can solve approximate 3-sided rectangle stabbing counting (ARSC) problem in $O(\log \log(n/\varepsilon k))$ time.*

3 A General Technique For Approximate Counting Query

Problem: $S = \{o_1, o_2, \dots, o_n\}$ is a set of n geometric objects. Let \mathcal{C} be the set of unique colors in S . Given a query object q , if k is the number of colors which have at least one object intersecting q , then report a value τ lying in the range $[(1 - \varepsilon)k, (1 + \varepsilon)k]$. In this section we present a general technique for efficiently solving any approximate counting query.

Assume that we have a corresponding *colored reporting structure* which reports all the colors which have at least one input object intersecting the query object. We present an efficient structure

for approximate counting by using colored reporting structure as a subroutine. Our results are summarized in the following two theorems.

Theorem 3. (Colored-GIQ) *Consider a colored-GIQ problem. Let $S(m)$ be the space occupied by the colored reporting structure when built on m objects; and let $O(Q_1(m) + tQ_2(m))$ be the query time to report t colors intersecting a query object. Then the approximate colored counting query can be solved using a data structure of size $O(S(n)\varepsilon^{-1} \log n)$ and in query time $O((Q_1(n) + \varepsilon^{-2} \log n \cdot Q_2(n)) \cdot \log(\varepsilon^{-1} \log n))$. The answer is correct with high probability.*

Theorem 4. (Standard GIQ) *Consider a standard GIQ problem. Let $S(m)$ be the space occupied by the reporting structure when built on m objects; and let $O(Q_1(m) + t)$ be the query time to report t objects intersecting a query object. Assume that the term $S(m)$ is geometrically converging. Then the approximate counting query can be solved:*

1. *using a data structure of size $O(S(n))$ and in query time $O((\varepsilon^{-1} \log n) [Q_1(\frac{n}{k}(\varepsilon^{-2} \log n)) + (\varepsilon^{-2} \log n)])$. The query algorithm is sensitive to the value of k .*
2. *using a data structure of size $O(S(n))$ and in query time $O([Q_1(n) + \varepsilon^{-2} \log n] \cdot \log(\varepsilon^{-1} \log n))$.*

The answer is correct with high probability.

3.1 Proof of Theorem 3

The proof of Theorem 3 is broken into two cases: In Subsubsection 3.1.1 we handle the case where k is small, and then in Subsubsection 3.1.2 and 3.1.3 we handle the case where k is large.

3.1.1 Handling small values of k

Based on the set S we build a colored reporting structure. Given a query object q , we query the colored reporting structure to keep reporting the colors in $S \cap q$ till one of the following event happens: either all the colors in $S \cap q$ have been reported or $\varepsilon^{-2} \log n + 1$ colors in $S \cap q$ have been reported. If the first event happens, then we have succeeded in obtaining the exact value of k . On the other hand, if the second event happens, then we can conclude that $k = \Omega(\varepsilon^{-2} \log n)$ (i.e., k is large). This query algorithm takes $O(Q_1(n) + \varepsilon^{-2} \log n \cdot Q_2(n))$ time.

3.1.2 Decision query

From now on we can safely assume that $k = |\mathcal{C} \cap q| = \Omega(\varepsilon^{-2} \log n)$. We start off by solving a decision problem: Given a number $z = \Omega(\varepsilon^{-2} \log n)$, is $|\mathcal{C} \cap q| \geq z$ or $|\mathcal{C} \cap q| < z$? The data structure is allowed to make a mistake when $|\mathcal{C} \cap q| \in [(1 - \varepsilon)z, (1 + \varepsilon)z]$. We will prove the following lemma.

Lemma 1. *The decision query can be solved using a data structure of size $O(S(n))$ and in query time $O(Q_1(n) + \varepsilon^{-2} \log n \cdot Q_2(n))$. The answer is correct with high probability.*

Data structure: A few words on the intuition behind our solution. Suppose each color in \mathcal{C} is sampled with probability $\approx (\log n)/z$. For a given query q , if $k < z$ (resp. $k > z$), then the expected number of colors from $\mathcal{C} \cap q$ sampled will be less than $\log n$ (resp. greater than $\log n$). This intuition is converted into an algorithm.

Set $M = (c_1 \varepsilon^{-2} \log n)/z$, where c_1 is a suitably large constant. Now a random sample of set \mathcal{C} is obtained as follows: each color in \mathcal{C} is independently picked with probability M . Now construct

a set S_M which consists of all the objects of S whose color got picked in the sample. A colored reporting structure is built based on the set S_M .

Query algorithm: Given a query object q , we query the colored reporting structure to keep reporting the colors in $S_M \cap q$ till one of the following event happens: either all the colors in $S_M \cap q$ have been reported or $c_1 \varepsilon^{-2} \log n + 1$ colors in $S_M \cap q$ have been reported. If the first event happens, then we claim that $|\mathcal{C} \cap q| \leq z$. On the other hand, if the second event happens, then we can conclude that $|\mathcal{C} \cap q| > z$. This query algorithm takes $O(Q_1(n) + \varepsilon^{-2} \log n \cdot Q_2(n))$ time.

Correctness: Now we show that with high probability the query algorithm returns the correct answer. For each of the k colors of S which intersect q define an indicator variable X_i . Set $X_i = 1$ if the corresponding color has objects in S_M ; otherwise set $X_i = 0$. Now define $Y = \sum_{i=1}^k X_i$. Then $E[Y] = k \cdot M$. Crucially Y is nothing but $|S_M \cap q|$.

When $k = z$, then $E[Y] = z \cdot M = c_1 \varepsilon^{-2} \log n$. The next lemma shows that the probability of the query algorithm reporting a wrong answer is small whenever $k = |\mathcal{C} \cap q| \geq (1 - \varepsilon)z$ or $k = |\mathcal{C} \cap q| < (1 + \varepsilon)z$.

Lemma 2. *When $k \leq (1 - \varepsilon)z$, then the probability of the event that $|S_M \cap q| > c_1 \varepsilon^{-2} \log n$ is small. Formally,*

$$\Pr[Y > zM \mid k \leq (1 - \varepsilon)z] \leq \frac{1}{n^{\Omega(1)}}.$$

Therefore, with high probability the query algorithm will claim that $k = |\mathcal{C} \cap q| \leq z$. Similarly,

$$\Pr[Y \leq zM \mid k \geq (1 + \varepsilon)z] \leq \frac{1}{n^{\Omega(1)}}$$

Proof. We only prove the first fact here. The proof for the second fact is similar. We will divide the proof into two cases based on the value of ε .

Case 1, $\varepsilon \in (0, 1/2]$: Let

$$\alpha = \Pr[Y > zM \mid k \leq (1 - \varepsilon)z]$$

The value α is maximized when $k = |\mathcal{C} \cap q| = (1 - \varepsilon)z$. Therefore,

$$\alpha \leq \Pr[Y > zM \mid k = (1 - \varepsilon)z]$$

In this case, $E[Y] = kM = (1 - \varepsilon)zM$. Therefore,

$$\begin{aligned} \alpha &\leq \Pr[Y > zM] = \Pr\left[Y > \frac{1}{1 - \varepsilon} E[Y]\right] \leq \Pr[Y > (1 + \varepsilon)E[Y]] \\ &\leq \exp\left(-\frac{\varepsilon^2 E[Y]}{4}\right) \quad \text{By Chernoff bound} \\ &= \exp\left(-\varepsilon^2 (1 - \varepsilon)z \left(\frac{c_1 \varepsilon^{-2} \log n}{4z}\right)\right) = \exp\left(-c_1 (1 - \varepsilon) \frac{\log n}{4}\right) \leq \exp\left(-\frac{c_1}{8} \log n\right) \quad \text{since } \varepsilon \leq 1/2 \\ &\leq \frac{1}{n^{\Omega(1)}} \end{aligned}$$

Case 2, $\varepsilon \in (1/2, 1]$: In this case we ignore the value of ε and set a new variable $\varepsilon_{new} \leftarrow 1/2$. The entire solution is built assuming that the error parameter is ε_{new} . Since $\varepsilon_{new} < \varepsilon$, clearly the error produced by this data structure will be within the tolerable limits. Also, observe that $\frac{1}{\varepsilon_{new}} \leq \frac{2}{\varepsilon}$. Therefore, the space and the query time bounds are also not affected. \square

3.1.3 Handling large values of k

Recall that we only have to handle $k = \Omega(\varepsilon^{-2} \log n)$. For the values $z_i = c_1(\varepsilon^{-2} \log n)(1 + \varepsilon)^i$, for $i = 1, 2, 3, \dots, W = O(\varepsilon^{-1} \log n)$, we build a data structure \mathcal{D}_i using Lemma 1. The overall space occupied will be $O(S(n)\varepsilon^{-1} \log n)$.

For a moment, assume that we query all the data structures $\mathcal{D}_1, \dots, \mathcal{D}_W$. Then with high probability, we will see a sequence of structures \mathcal{D}_j for $j \in [1, i]$ claiming $|\mathcal{C} \cap q| > z_j$, followed by a sequence of structures $\mathcal{D}_{i+1}, \dots, \mathcal{D}_W$ claiming $|\mathcal{C} \cap q| \leq z_j$. Then we shall report $\tau \leftarrow z_i$ as the answer to the approximate colored counting query. A simple calculation reveals that $\tau \in [(1 - \varepsilon)k, k]$. We can perform binary search on $\mathcal{D}_1, \dots, \mathcal{D}_W$ to efficiently find the index i . The overall query time will be $O((Q_1(n) + \varepsilon^{-2} \log n \cdot Q_2(n)) \cdot \log(\varepsilon^{-1} \log n))$. This finishes the proof of Theorem 3.

3.2 Proof of Theorem 4

A standard GIQ problem is a special case of its corresponding colored-GIQ problem; assign each object in the standard GIQ problem a unique color. The data structure built in Subsection 3.1 for a colored GIQ problem will be used to solve the standard GIQ problem as well. Interestingly, in this special setting one can obtain a sub-linear bound on the size of set S_M , which is not possible in the general case. For a fixed value of M , the size of S_M is $O(nM)$. Therefore, Lemma 1 can be refined as follows.

Lemma 3. *The decision query can be solved using a data structure of size $O(S(nM))$ and in query time $O(Q_1(nM) + \varepsilon^{-2} \log n)$, where $M = (c_1 \varepsilon^{-2} \log n)/z$. The answer is correct with high probability.*

Now we will bound the space occupied by all the decision data structures $\mathcal{D}_1, \dots, \mathcal{D}_W$. Since the term $S(\cdot)$ is geometrically converging, the total space will be

$$\sum_{i=1}^W O\left(S\left(\frac{n\varepsilon^{-2} \log n}{z_i}\right)\right) = O\left(S\left(\frac{n\varepsilon^{-2} \log n}{z_1}\right)\right) = O\left(S\left(\frac{n}{1+\varepsilon}\right)\right) = O(S(n)).$$

We propose two different query algorithms:

(1) Query the data structures in the order $\mathcal{D}_W, \mathcal{D}_{W-1}, \mathcal{D}_{W-2}, \dots, \mathcal{D}_i$ to find the index i . Since \mathcal{D}_i is the largest-size data structure queried, the overall query time will be bounded by

$$\begin{aligned} & \leq \overbrace{\left(\frac{\# \text{ structures queried}}{W-i}\right)}^{\text{decision query on } \mathcal{D}_i} \cdot O\left(Q_1\left(\frac{n\varepsilon^{-2} \log n}{z_i}\right) + \varepsilon^{-2} \log n\right) \\ & \leq (\varepsilon^{-1} \log n) \cdot O\left(Q_1\left(\frac{n}{k}(\varepsilon^{-2} \log n)\right) + (\varepsilon^{-2} \log n)\right) \quad \text{since } z_i \geq (1 - \varepsilon)k \geq k/2 \end{aligned}$$

This proves the first bullet of Theorem 4.

(2) Perform binary search on $\mathcal{D}_1, \dots, \mathcal{D}_W$ to efficiently find the index i . The overall query time will be $O((Q_1(n) + \varepsilon^{-2} \log n) \cdot \log(\varepsilon^{-1} \log n))$. This proves the second bullet of Theorem 4.

3.3 Applications

We present two applications of our technique. The first one is the approximate halfspace range counting in $\mathbb{R}^d, d \geq 4$. Afshani and Chan [4] presented an $O(n)$ space structure for halfspace range reporting in $\mathbb{R}^d, d \geq 4$ which can solve the query in $\tilde{O}(n^{1-1/\lfloor d/2 \rfloor} + k)$ time. Applying the first bullet of Theorem 4, we obtain the following result.

Corollary 1. *There is a data structure of size $O(n)$ which can solve the approximate halfspace range counting in \mathbb{R}^d , $d \geq 4$ in $\tilde{O}((n/k)^{1-1/\lfloor d/2 \rfloor})$ time. The answer is correct with high probability.*

As an application of Theorem 3, consider the approximate orthogonal colored range counting. The proof of the corollary can be found in Section 8 of the appendix.

Corollary 2. *In the orthogonal colored range counting the input set S is n colored points in d -dimensional space and the query q is a d -dimensional rectangle. There is a data structure of size $O(\varepsilon^{-1} n \log^{d+1} n)$ which can answer the approximate counting query in $O(\varepsilon^{-2} \log^{d+1} n \cdot \log(\varepsilon^{-1} \log n))$ time. The answer is correct with high probability.*

3.4 Enhancement of Aronov and Har-Peled's technique?

We claim that the approximate counting technique proposed in this section is an enhancement of Aronov and Har-Peled's (AH's) approximate counting technique [15]. Some reasons for our belief are the following:

- Unlike AH's technique which is limited to standard GIQ problems, our technique can be applied to efficiently solve colored-GIQ problems (Theorem 3).
- Unlike AH's technique, the first bullet of Theorem 4 leads to an algorithm whose query time is inversely proportional to the value of k , which is a desirable property to have.
- Unlike AH's structure, the space occupied by our structure in Theorem 4 is independent of ε .

4 Approximate Orthogonal Colored Range Counting

In this section we study the 3-sided approximate orthogonal colored range counting (*AOCRC*) problem in \mathbb{R}^2 . The input set S is a set of colored points in \mathbb{R}^2 and the query q is a 3-sided rectangle in \mathbb{R}^2 . The following result is obtained in this section.

Theorem 5. *There exists a data structure of size $O(\varepsilon^{-2} n)$ which can solve 3-sided *AOCRC* problem in $(\varepsilon^{-2} \log n)$ worst-case time. With constant positive probability (such as 99/100 or 999/1000) τ lies in the range $[(1 - \varepsilon)k, (1 + \varepsilon)k]$.*

4.1 Reduction to 5-sided rectangle stabbing in \mathbb{R}^3

In this subsection we present a reduction of 3-sided *AOCRC* problem to 5-sided approximate rectangle stabbing counting (*ARSC*) problem in \mathbb{R}^3 . Let S be a set of n colored points lying in \mathbb{R}^2 . Let $S_c \subseteq S$ be the set of points of color c . For each color c which has at least one point inside $q = [x_1, x_2] \times [y_1, \infty)$, the objective is to identify the topmost point (in terms of y -coordinate) among $S_c \cap q$. Consider a point $p(p_x, p_y) \in S_c$. Starting from the x -coordinate value p_x , we walk to the left (resp. right) along the x -axis till we find the first point $p^l(p_x^l, p_y^l) \in S_c$ (resp. $p^r(p_x^r, p_y^r) \in S_c$) which has a higher y -coordinate value than p . (Conceptually imagine two dummy points at $(+\infty, +\infty)$ and $(-\infty, +\infty)$ to ensure that p^l and p^r always exist). Now we make the following important observation.

Lemma 4. *A point $p \in S_c$ will be the topmost point in $S_c \cap q$ iff (1) p lies inside q , and (2) p_r and p_l do not lie inside q . In other words, $p \in S_c$ will be the topmost point in $S_c \cap q$ iff $(x_1, x_2, y_1) \in [p_x^l, p_x] \times [p_x, p_x^r] \times (-\infty, p_y]$.*

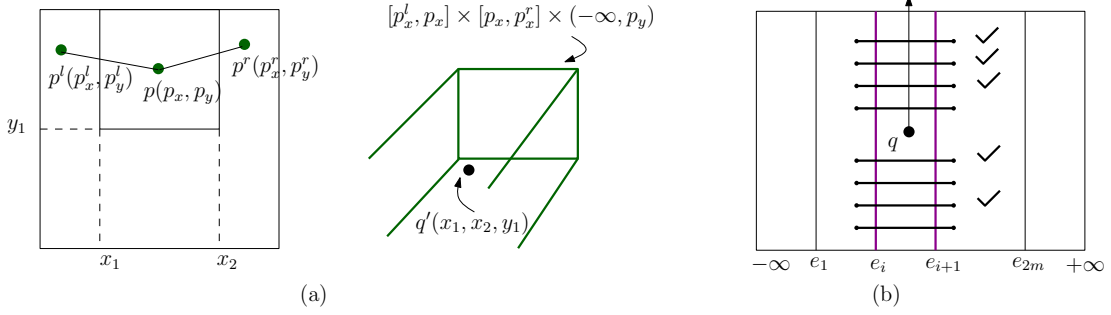


Figure 2: (a) Reduction from 3-sided approximate orthogonal colored range counting (AOCRC) problem in \mathbb{R}^2 to 5-sided approximate rectangle stabbing counting (ARSC) problem in \mathbb{R}^3 . (b) Set R^{es} for the elementary segment $es = (e_i, e_{i+1})$. For $\varepsilon = 1/2$, the ticked (\checkmark) entries are the ones which will be stored in the sketch. To avoid clutter in the figure, only the top segment $[y_1, y_2]$ of the rectangles of R^{es} is shown in the figure.

Figure 2(a) is an illustration of the above observation. Based on the above observation, we perform the following transformation: Each point $p \in S$ is transformed into a 5-sided rectangle $[p_x^l, p_x] \times [p_x, p_x^r] \times (-\infty, p_y]$. The query rectangle $q = [x_1, x_2] \times [y_1, \infty)$ is transformed into a point $q'(x_1, x_2, y_1) \in \mathbb{R}^3$. Now we can observe that (i) If a color c has at least one point inside q , then exactly one of its transformed rectangle will contain q' , and (ii) If a color c has no point inside q , then none of its transformed rectangles will contain q' . Therefore, an approximate counting query on the 5-sided rectangles will answer the 3-sided AOCRC problem.

4.2 When $k \in [C\varepsilon^{-2} \log^4 n, n]$

Using the above reduction, in this subsection we study the 5-sided ARSC problem in \mathbb{R}^3 and prove the following theorem.

Theorem 6. *Suppose $k \geq C\varepsilon^{-2} \log^4 n$, where C is a suitably large constant. Then there exists a data structure of size $O(\varepsilon^{-2}n)$ which can solve 5-sided ARSC problem in $O(\log n)$ worst-case time.*

This implies a data structure of size $O(\varepsilon^{-2}n)$ which can solve 3-sided AOCRC problem in $O(\log n)$ worst-case time.

A brief overview of the proof of Theorem 6. The idea of using random sampling for answering an approximate range counting query has been used in the past in [15, 20, 26] which deal with non-colored objects. Random sampling helps reduce the size of the input set and thus, allows us to use a slightly space-inferior data structure. Theorem 7 presents the slightly space-inferior but crucially query time optimal structure for solving 5-sided ARSC problem. Then the random sampling technique of Lemma 5 is applied along with Theorem 7 to obtain a space optimal and query time optimal solution when $k \geq C\varepsilon^{-2} \log^4 n$.

Theorem 7. *Let R be a set of m 5-sided rectangles in \mathbb{R}^3 . Then there exists a data structure of size $O(\varepsilon^{-2}m \log^3 m)$ which can solve 5-sided ARSC problem in $O(\log m)$ worst-case time.*

Intuition behind proof of Theorem 7: The formal proof of Theorem 7 can be found in Section 9. We only provide some intuition behind the proof. Consider a simpler setting where the set R is m 3-sided rectangles of the form $[y_1, y_2] \times (-\infty, z]$ in \mathbb{R}^2 . Project these rectangles onto the y -axis. let $E_y = (e_1, e_2, \dots, e_{2m})$ be the sorted sequence (in increasing order of y -coordinate) of the

endpoints of these projected intervals. We divide the y -axis into *elementary segments* $(-\infty, e_1)$, $[e_1, e_1]$, (e_1, e_2) , $[e_2, e_2]$, (e_2, e_3) , \dots , (e_{2m-1}, e_{2m}) , $[e_{2m}, e_{2m}]$, (e_{2m}, ∞) . For any elementary segment, say es , let R^{es} be the set of rectangles completely crossing the segment es . If the query point q lies in es , to answer the approximate counting query it is enough to store a *sketch* of R^{es} . See Figure 2(b). The size of the sketch of R^{es} is only $O(\varepsilon^{-1} \log m)$. The total size of all the sketches will be $O(\varepsilon^{-1} m \log m)$.

To handle the general case of 5-sided rectangles, we use ideas such as (a) constructing an “external-memory style” segment tree [14] with fanout $\varepsilon^{-1} \log m$, and (b) fractional cascading to efficiently query the $O\left(\frac{\log m}{\log(\varepsilon^{-1} \log m)}\right)$ sketches in the segment tree. \square

Lemma 5. *For a particular GIQ problem, let S be a set of n objects in \mathbb{R}^d . We require the following two conditions to hold:*

1. *The number of combinatorially different query objects on the set S is bounded by $O(n^{c_1})$, where c_1 is a constant independent of n and ε .*
2. *The query object q has to be δ -heavy. A query object q is called δ -heavy if $k \geq \delta(C\varepsilon^{-2} \log n)$.*

Then there exists a set $R \subset S$ of size $O(n/\delta)$ such that for any δ -heavy query, $(|R \cap q| \cdot \delta) \in [(1 - \varepsilon)k, (1 + \varepsilon)k]$.

Proof. Construct a random sample R where each object of S is picked with probability $1/\delta$. Therefore, the expected size of R is n/δ (if the size of R exceeds $O(n/\delta)$ then we re-sample till we get the desired size). For a given query q , $E[|R \cap q|] = |S \cap q|/\delta = k/\delta$. Therefore, by Chernoff bound [37] we observe that

$$\Pr \left[\left| |R \cap q| - \frac{k}{\delta} \right| > \varepsilon \frac{k}{\delta} \right] \leq e^{-\Omega(\varepsilon^2(k/\delta))} \leq e^{-\Omega(\varepsilon^2(C\varepsilon^{-2} \log n))} \leq e^{-\Omega(C \log n)} = n^{-\Omega(C)} \leq o(1/n^C)$$

We will pick a C such that $C > c_1$. Then observe that, as there are only $O(n^{c_1})$ number of combinatorially different query objects on the set S , by union bound it follows that there exists a subset $R \subset S$ of size $O(n/\delta)$ such that for any δ -heavy query range $|k - |R \cap q| \cdot \delta| \leq \varepsilon k$. \square

Final Structure(Combining Theorem 7 and Lemma 5): Let S be a set of n 5-sided rectangles in \mathbb{R}^3 . The number of combinatorially different query points on S is bounded by $O(n^3)$. We set $\delta \leftarrow \log^3 n$ and define a new parameter $\varepsilon' \leftarrow \varepsilon/4$. Now we apply Lemma 5 to obtain a set R of size $O(n/\log^3 n)$. Based on the set R and with error parameter ε' , we build the data structure of Theorem 7. Given a δ -heavy query on S , we query the data structure built on R . Let τ_R be the value returned. Then we report $\tau_R \log^3 n$ as the answer to the 5-sided ARSC problem on S .

Analysis: Since $|R| = O(n/\log^3 n)$, by Theorem 7 the space occupied by our data structure will be $O(\varepsilon^{-2}n)$. By Theorem 7, the query time will be $O(\log n)$. Next, we will prove that $(1 - \varepsilon)k \leq \tau_R \log^3 n \leq (1 + \varepsilon)k$.

If we knew the exact count of $|R \cap q|$, then from Lemma 5 we can infer that:

$$(1 - \varepsilon')k \leq |R \cap q| \log^3 n \leq (1 + \varepsilon')k \quad (1)$$

However, by using Theorem 7 we only get the following approximation of $|R \cap q|$:

$$(1 - \varepsilon')|R \cap q| \leq \tau_R \leq (1 + \varepsilon')|R \cap q| \quad (2)$$

Combining the above two equations, we get the following:

$$\begin{aligned}
(1 - \varepsilon')^2 k &\leq (1 - \varepsilon') |R \cap q| \log^3 n \leq \tau_R \log^3 n \leq (1 + \varepsilon') |R \cap q| \log^3 n \leq (1 + \varepsilon')^2 k \\
\implies (1 + \varepsilon'^2 - 2\varepsilon') k &\leq \tau_R \log^3 n \leq (1 + \varepsilon'^2 + 2\varepsilon') k \\
\implies (1 - \varepsilon) k &\leq \tau_R \log^3 n \leq (1 + \varepsilon) k \quad \text{where } \varepsilon = 4\varepsilon'
\end{aligned}$$

This finishes the proof of Theorem 6.

4.3 When $k \in [0, C\varepsilon^{-2} \log^4 n]$

To handle this case, we will directly work with 3-sided *AOCRC* and not use the reduction to 5-sided *ARSC*. We state a lemma which is similar to Lemma 2, in the sense that random sampling on colors will be performed.

Lemma 6. *Let S be a set of n colored points in \mathbb{R}^2 and let \mathcal{C} be the set of unique colors in S . A 3-sided query rectangle q is called δ -heavy if $k \geq C\varepsilon^{-2}\delta$, where $k = |\mathcal{C} \cap q|$ is the number of unique colors in $S \cap q$. Then there exists a set $R \subset \mathcal{C}$ of size such that with certain constant positive probability $|k - |R \cap q|| \leq \varepsilon k$, for any δ -heavy query. Here $R \cap q$ denotes the set of colors in R which have at least one point inside q .*

Proof. Construct a random sample R where each color in \mathcal{C} is picked independently with probability $1/\delta$. For a given query q , $E[|R \cap q|] = |\mathcal{C} \cap q|/\delta = k/\delta$. Therefore, by Chernoff bound [37] we observe that

$$\Pr \left[\left| |R \cap q| - \frac{k}{\delta} \right| > \varepsilon \frac{k}{\delta} \right] \leq e^{-\Omega(\varepsilon^2(k/\delta))} \leq e^{-\Omega(\varepsilon^2(C\varepsilon^{-2}))} < e^{-\Omega(C)} < 1$$

by assuming sufficiently large C . □

Now we are ready to present the solution for all the sub-cases:

(1) *When $k \in [0, C\varepsilon^{-2} \log n]$:* Shi and Jaja [42] present a structure of size $O(n)$ which can answer a 3-sided orthogonal colored range reporting query in $O(\log n + k)$ time. Build this structure on all the points of set S . Given a query rectangle q , we query the structure to keep reporting the colors in $S \cap q$ till one of the following event happens: either all the colors in $S \cap q$ have been reported or $C\varepsilon^{-2} \log n + 1$ colors in $S \cap q$ have been reported. If the first event happens, then we have succeeded in obtaining the exact value of k in $O(\varepsilon^{-2} \log n)$ time. On the other hand, if the second event happens, then we can conclude that $k > C\varepsilon^{-2} \log n$.

(2) *When $k \in [C\varepsilon^{-2} \log n, C\varepsilon^{-2} \log^2 n]$:* We shall now use Lemma 6. We set $\delta = \log n$. Then as per the requirement of Lemma 6, the query q is indeed $C\varepsilon^{-2} \log n$ -heavy. To apply Lemma 6 efficiently, we need a structure for computing $|R \cap q|$. For this we build the reporting structure of Shi and Jaja [42] on all the points of S whose color is in set R . Given a query rectangle q , we query this structure to keep reporting all the colors in $R \cap q$ till one of the following event happens: either all the colors in $R \cap q$ have been reported or $2C\varepsilon^{-2} \log n + 1$ colors in $R \cap q$ have been reported. If the first event happens, then we have succeeded in obtaining the exact value of $|R \cap q|$ in $O(\varepsilon^{-2} \log n)$ time. We report $|R \cap q| \cdot \delta$ as the answer. On the other hand, if the second event happens, then we conclude that with certain constant positive probability:

$$(1 + \varepsilon)k \geq |R \cap q| \cdot \delta > 2C\varepsilon^{-2} \log^2 n \implies k > C\varepsilon^{-2} \log^2 n$$

(3) When $k \in [C\varepsilon^{-2} \log^2 n, C\varepsilon^{-2} \log^3 n]$ and $k \in [C\varepsilon^{-2} \log^3 n, C\varepsilon^{-2} \log^4 n]$: They are handled similar to case (2), with the only difference that query q is $C\varepsilon^{-2} \log^2 n$ -heavy and $C\varepsilon^{-2} \log^3 n$ -heavy, respectively.

Lemma 7. *When $k \leq C\varepsilon^{-2} \log^4 n$, there exists a data structure of size $O(n)$ which can solve the 3-sided AOCRC problem in $O(\varepsilon^{-2} \log n)$ worst-case time. With constant probability (such as 99/100 or 999/1000) the value of τ lies in the range $[(1 - \varepsilon)k, (1 + \varepsilon)k]$.*

APPENDIX for Section 2

5 Correctness of the query algorithm in section 2.2

Given a query interval $q = (-\infty, q_x]$, let \mathcal{I}_i be the interval containing q_x and $L(i)$ be the predecessor of q_x . We look at various ranges of i and handle each of them separately:

- (i) When $i < \lceil 1/\varepsilon' \rceil$: In this case we report the exact value of k . Therefore, $\tau = i = k$.
- (ii) When $i = \lceil 1/\varepsilon' \rceil$: In this case the predecessor of q_x , $L(\lceil 1/\varepsilon' \rceil)$, has rank $\lceil 1/\varepsilon' \rceil$. The successor of q_x , $L(\lceil 1/\varepsilon' \rceil + 1)$, has a rank less than $\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^2 \rceil$. We report $\tau = \lceil 1/\varepsilon' \rceil$. Therefore,

$$\begin{aligned}
 k \in \left[\lceil 1/\varepsilon' \rceil, \left\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^2 \right\rceil \right) &\implies k \in \left[\lceil 1/\varepsilon' \rceil, 1 + \frac{1}{\varepsilon'}(1 + \varepsilon')^2 \right) \implies k \in \left[\lceil 1/\varepsilon' \rceil, 1 + \frac{1}{\varepsilon'}(1 + 3\varepsilon') \right) \\
 \implies k \in \left[\lceil 1/\varepsilon' \rceil, \frac{1}{\varepsilon'}(1 + C\varepsilon') \right) &\implies k \in [\lceil 1/\varepsilon' \rceil, \lceil 1/\varepsilon' \rceil (1 + C\varepsilon')] \implies k \in [\tau, \tau(1 + C\varepsilon')] \\
 \implies \tau \in ((1 - C\varepsilon')k, k] &\implies \tau \in ((1 - \varepsilon)k, (1 + \varepsilon)k) \text{ by setting } \varepsilon = C\varepsilon'
 \end{aligned}$$

- (iii) When $i \in [\lceil 1/\varepsilon' \rceil + 1, \lceil 1/\varepsilon' \rceil + j]$: In this case the predecessor of q_x , $L(i)$, has rank greater than $\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2(i-1)} \rceil$. The successor of q_x , $L(i + 1)$, has rank less than $\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2(i+1)} \rceil$.

$$\begin{aligned}
 k &\in \left(\left\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2(i-1)} \right\rceil, \left\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2(i+1)} \right\rceil \right) \\
 \implies k &\in \left(\frac{1}{1 + \varepsilon'} \left\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i-1} \right\rceil, 1 + \frac{1}{\varepsilon'}(1 + \varepsilon')^{2(i+1)} \right) \\
 \implies k &\in \left(\frac{1}{1 + \varepsilon'} \left\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i-1} \right\rceil, 1 + \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i-1}(1 + \varepsilon')^3 \right) \\
 \implies k &\in \left(\frac{1}{1 + \varepsilon'} \left\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i-1} \right\rceil, (1 + \varepsilon')^{2i-1} + \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i-1}(1 + \varepsilon')^3 \right) \\
 \implies k &\in \left(\frac{1}{1 + \varepsilon'} \left\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i-1} \right\rceil, \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i-1}(\varepsilon' + (1 + \varepsilon')^3) \right) \\
 \implies k &\in \left(\frac{1}{1 + \varepsilon'} \left\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i-1} \right\rceil, \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i-1}(1 + C\varepsilon') \right) \\
 \implies k &\in \left(\tau \left(\frac{1}{1 + \varepsilon'} \right), \tau(1 + C\varepsilon') \right) \\
 \implies \tau &\in ((1 - C\varepsilon')k, k(1 + \varepsilon')) \\
 \implies \tau &\in ((1 - \varepsilon)k, (1 + \varepsilon)k) \text{ by setting } \varepsilon = C\varepsilon'
 \end{aligned}$$

- (iv) When $i = \lceil 1/\varepsilon' \rceil + j + 1$: In this case the predecessor of q_x , $L(i)$, has rank greater than $\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2j} \rceil$. Recall that j is the largest integer such that $m_t \geq \lceil \frac{1}{\varepsilon}(1 + \varepsilon)^{2j+1} \rceil$. This implies that $m_t < \lceil \frac{1}{\varepsilon}(1 + \varepsilon)^{2(j+1)+1} \rceil$. Therefore, we can conclude that

$$\begin{aligned}
& k \in \left(\left\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2j} \right\rceil, \left\lceil \frac{1}{\varepsilon}(1 + \varepsilon)^{2(j+1)+1} \right\rceil \right) \\
\implies & k \in \left(\tau \left(\frac{1}{1 + \varepsilon'} \right), \tau(1 + C\varepsilon') \right) \text{ by similar calculations as done above} \\
\implies & \tau \in ((1 - \varepsilon)k, (1 + \varepsilon)k) \text{ by setting } \varepsilon = C\varepsilon'
\end{aligned}$$

6 Upper bound on the number of changes to list L

Let the $(t + 1)$ -th update be *deletion* of a point from P , i.e., $m_{t+1} = m_t - 1$. The invariants are fixed as follows:

(1) *Updating existing entries in L* : Some of the entries in L might start violating Invariant 2 or Invariant 4. To fix this, each violating entry is set to its median position.

(2) *Deleting the last entry in L* : Again we look at three different scenarios. First, when $m_{t+1} < \lceil 1/\varepsilon' \rceil$. Then we delete the last entry in the list L . Second, when $m_{t+1} = \lceil 1/\varepsilon' \rceil$. Then $m_t = \lceil 1/\varepsilon' \rceil + 1 = \lceil \frac{1}{\varepsilon'}(1 + \varepsilon') \rceil$. Then we delete the last entry (i.e. $L(\lceil 1/\varepsilon' \rceil + 1)$). These two cases fix the violation of Invariant 1 and 2.

Finally, we look at the case where $m_{t+1} > \lceil 1/\varepsilon' \rceil$. Recall that j is the largest integer such that $m_t \geq \lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2j+1} \rceil$. After the deletion of a point from P , it might happen that $m_{t+1} < \lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2j+1} \rceil$. Then we delete the last entry $L(\lceil 1/\varepsilon' \rceil + j + 1)$ from the list L . This fixes the violation of Invariant 3 and 4.

Now we prove that the total number of changes made to the list L is bounded by $O(\varepsilon^{-1}m)$. Every update operation to set P can lead to either creation of a new entry in L or deletion of the last entry in L . Luckily these types of changes are bounded by $O(m)$. Now we upper bound the number of changes happening to the list L because of its entries violating either Invariant 2 or Invariant 4. We divide them into two cases:

Case a: Consider the first c_1/ε' entries of L , for a sufficiently large constant c_1 . We make the trivial observation that every update to P could lead to these entries violating Invariant 2 or Invariant 4. Therefore, the total number of changes made to the first c_1/ε' entries is bounded by $O(mc_1/\varepsilon') = O(m/\varepsilon)$.

Case b: Now consider the remaining entries of L . Once an entry $L(\lceil 1/\varepsilon' \rceil + i)$ has been set to its median position, then:

1. at least $\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i-1} \rceil - \lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2(i-1)} \rceil$ deletion of points of P has to happen, or
2. at least $\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i} \rceil - \lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i-1} \rceil$ insertion of points of P has to happen

before $L(\lceil 1/\varepsilon' \rceil + i)$ starts violating Invariant 4. Now we prove the following mathematical claim.

Observation 1. $\lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2i-1} \rceil - \lceil \frac{1}{\varepsilon'}(1 + \varepsilon')^{2(i-1)} \rceil \geq (1/2)e^{(i-1)\varepsilon'}$

Proof.

$$\begin{aligned}
& \left(\left\lceil \frac{1}{\varepsilon'} (1 + \varepsilon')^{2i-1} \right\rceil \right) - \left(\left\lceil \frac{1}{\varepsilon'} (1 + \varepsilon')^{2(i-1)} \right\rceil \right) \geq \left(\frac{1}{\varepsilon'} (1 + \varepsilon')^{2i-1} \right) - \left(\frac{1}{\varepsilon'} (1 + \varepsilon')^{2(i-1)} + 1 \right) \\
& = \frac{1}{\varepsilon'} (1 + \varepsilon')^{2i-2} \cdot [(1 + \varepsilon') - 1] - 1 = (1 + \varepsilon')^{2i-2} - 1 = \left[(1 + \varepsilon')^{2/\varepsilon'} \right]^{(i-1)\varepsilon'} - 1 \\
& \geq \exp((i-1)\varepsilon') - 1 \quad \text{since } 1 + x \geq e^{x/2}, 0 \leq x \leq 2 \\
& \geq (1/2) \exp((i-1)\varepsilon') \quad \text{since } i = \Omega(c_1/\varepsilon')
\end{aligned}$$

□

Therefore, during the m updates, the number of times the entry $L(\lceil 1/\varepsilon' \rceil + i)$ can violate Invariant 4 because of deletion of at least $\lceil \frac{1}{\varepsilon'} (1 + \varepsilon')^{2i-1} \rceil - \lceil \frac{1}{\varepsilon'} (1 + \varepsilon')^{2(i-1)} \rceil$ points of P is

$$\leq \frac{2m}{e^{(i-1)\varepsilon'}}$$

Summing up the above quantity for all the entries of L being considered, we get

$$\begin{aligned}
\# \text{ changes in } L & \leq \sum \left(\frac{2m}{e^{(i-1)\varepsilon'}} \right) \\
& \leq O(m/\varepsilon') \quad \text{a loose upper bound}
\end{aligned}$$

A similar calculation reveals that $O(m/\varepsilon')$ is the number of times entries in L violate Invariant 4 because of insertion of points of P . This finishes our proof.

7 Proof of Theorem 2

To obtain time bound of $O(\log \log(n/\varepsilon k))$, the time spent by the query algorithm has to be inversely proportional to the value of k . Therefore, our strategy would be to first check if k is large and if not, then progressively check for smaller values of k . When the value of k is guaranteed to be below $\sqrt{\frac{n}{\varepsilon}}$, then we can afford a query time of $O(\log \log(\varepsilon^{-1}n))$, since $O(\log \log(n/\varepsilon k)) = O(\log \log(\varepsilon^{-1}n))$ when $k < \sqrt{\frac{n}{\varepsilon}}$. In this section we will look at the case when $k \in [\sqrt{\frac{n}{\varepsilon}}, n]$ and prove the following result.

Lemma 8. *When $k \in [\sqrt{\frac{n}{\varepsilon}}, n]$, there exists a data structure of size $O(\varepsilon^{-1}n)$ which can answer ARSC problem in \mathbb{R}^2 in $O(\log \log(n/\varepsilon k))$ worst-case time.*

First, we consider the case where $\varepsilon \in (1/2, 1)$. In this case we ignore the value of ε and set a new variable $\varepsilon_{new} \leftarrow 1/2$. Now the data structure is built assuming the error parameter is ε_{new} . Since $\varepsilon_{new} < \varepsilon$, the error produced by the data structure will be within the tolerable limits. Plus, we have $\varepsilon \leq 2\varepsilon_{new}$; which implies $\frac{1}{\varepsilon_{new}} \leq \frac{2}{\varepsilon}$. Therefore, the space and the query time bounds are also not affected.

From now on we can safely assume that $\varepsilon \in (0, 1/2]$. We will present a routine **TestDepth** later in this subsection. **TestDepth** takes a parameter \hat{k} and depending on the value of k performs one of the following operations in $O(\log \log(n/\varepsilon \hat{k}))$ time:

1. If $k \geq \frac{\hat{k}}{1-\varepsilon}$, then return a value τ such that $\tau \in [(1-\varepsilon)k, (1+\varepsilon)k]$.
2. If $k < \hat{k}$, then return the value **shallow**.

3. If $\hat{k} \leq k < \frac{\hat{k}}{1-\varepsilon}$, then **TestDepth** is allowed to perform either step (1) or step (2).

We make a series of calls to **TestDepth** with various values of \hat{k} . The key idea is to make calls with triple exponentially decreasing values. Specifically, starting from $i = 0$, in the i -th call to **TestDepth** we set $\hat{k} = \frac{2n}{\varepsilon \cdot 2^{2^{2^i}}}$. If the returned value is **shallow**, then we perform an $(i+1)$ -th call to **TestDepth**. Otherwise, we have obtained a value τ s.t. $\tau \in [(1-\varepsilon)k, (1+\varepsilon)k]$ and we stop.

Analysis: To answer a query, let j be the number of calls to **TestDepth**. Then the total time taken by all the calls to **TestDepth** will be

$$\sum_{i=0}^j O\left(\log \log \left(\frac{n}{\varepsilon \cdot (2n/\varepsilon \cdot 2^{2^{2^i}})}\right)\right) = \sum_{i=0}^j O\left(\log \log 2^{2^{2^i}}\right) = \sum_{i=0}^j O(2^i) = O(2^j)$$

Since the $(j-1)$ -th call to **TestDepth** returned **shallow** (otherwise the j -th call would not have been made), we can conclude that

$$\begin{aligned} k < \frac{1}{1-\varepsilon} \left(\frac{2n}{\varepsilon \cdot 2^{2^{2^{j-1}}}} \right) &\implies 2^{2^{2^{j-1}}} < \frac{2}{\varepsilon(1-\varepsilon)}(n/k) \leq \frac{2}{\varepsilon^2}(n/k) \quad \text{since } \varepsilon \leq 1/2 \text{ implies } \frac{1}{1-\varepsilon} \leq \frac{1}{\varepsilon} \\ \implies 2^{2^j} &= O\left(\log \frac{1}{\varepsilon} + \log \frac{n}{k}\right) \end{aligned}$$

Therefore, the total time taken by all the calls to **TestDepth** is bounded by $O(\log \log \frac{1}{\varepsilon} + \log \log \frac{n}{k})$.

7.0.1 Implementation of TestDepth

Now we present the details of **TestDepth**. Assume that we need to build the data structure for some fixed value of the parameter \hat{k} . Construct an orthogonal grid $(4n/\varepsilon\hat{k}) \times (4n/\varepsilon\hat{k})$ such that each horizontal and vertical slab contains $\varepsilon\hat{k}/2$ sides of the rectangles in S . For each cell c in the grid we keep a value, say $val(c)$, which is the number of rectangles in S which completely cover the cell c . The space occupied by this structure is $O((n/\varepsilon\hat{k})^2)$.

Given a query point q , we can locate the cell c containing q in $O(\log(n/\varepsilon\hat{k}))$ time. If $val(c) < \hat{k}$, then return **shallow**. Otherwise, if $val(c) \geq \hat{k}$, then return $\tau = val(c)$.

We analyze the space occupied by **TestDepth**. For a fixed value of \hat{k} , the space occupied is $O((n/\varepsilon\hat{k})^2)$. Adding this up over all the values of \hat{k} with which **TestDepth** is called (i.e., $\hat{k} \leftarrow 2n/\varepsilon \cdot 2^{2^{2^0}}, 2n/\varepsilon \cdot 2^{2^{2^1}}, \dots, \sqrt{\frac{n}{\varepsilon}}$), leads to $O(\varepsilon^{-1}n)$.

We prove that **TestDepth** performs its operations correctly. For a cell c containing q , it is easy to observe that $val(c) \leq k \leq val(c) + \varepsilon\hat{k}$, since there can be at most $2(\varepsilon\hat{k}/2) = \varepsilon\hat{k}$ rectangles in S which can partially intersect cell c . If $k < \hat{k}$, then it implies that $val(c) \leq k < \hat{k}$. This implies that the algorithm will return **shallow**. If $k \geq \frac{\hat{k}}{1-\varepsilon}$, then we prove that $val(c) \geq \hat{k}$. Combining the facts that $k \geq \frac{\hat{k}}{1-\varepsilon}$ and $k \leq val(c) + \varepsilon\hat{k}$, we get $\frac{\hat{k}}{1-\varepsilon} \leq val(c) + \varepsilon\hat{k}$. Rearranging this equation we get $val(c) \geq \hat{k} \left(1 + \frac{\varepsilon^2}{1-\varepsilon}\right) \geq \hat{k}$. Since $val(c) \geq \hat{k}$, **TestDepth** will return a value $\tau = val(c)$.

Finally, we prove that when the query algorithm returns τ , it will be in the range $[(1-\varepsilon)k, (1+\varepsilon)k]$. Since $val(c) \leq k$, it is obvious that $\tau = val(c)$ will be $\leq (1+\varepsilon)k$. To prove the other direction, we combine the fact that $val(c) \geq \hat{k}$ and $k \leq val(c) + \varepsilon\hat{k}$, to get $k \leq val(c) + \varepsilon \cdot val(c) = \tau(1+\varepsilon)$.

Manipulating $k \leq \tau(1 + \varepsilon)$ we can get $\tau \geq (1 - \varepsilon)k$. This proves that the routine **TestDepth** performs its operations correctly.

The above discussion easily generalizes to higher dimensions. So, we state the following lemma without a proof to avoid repetition.

Lemma 9. *Let S be a set of n rectangles in \mathbb{R}^d . Given a query point q in \mathbb{R}^d and a parameter \hat{k} , the routine **TestDepth** can be implemented using $O((n/\varepsilon\hat{k})^d)$ space and it can perform its operations in $O(\log \log(n/\varepsilon\hat{k}))$ time.*

APPENDIX for Section 3

8 Proof of corollary 2

First consider the non-colored orthogonal range emptiness query in \mathbb{R}^d ($d \geq 2$). Using the standard range tree, this problem can be solved using $M(n) = O(n \log^{d-1} n)$ space and $f(n) = O(\log^{d-1} n)$ query time. Using this structure, the orthogonal colored range reporting problem in \mathbb{R}^d can be answered in $O(f(n) + \kappa f(n) \log n)$ query time using a structure of size $O(M(n) \log n)$. Here κ is the number of colors reported (see [25] for the details of this transformation). Therefore, the space occupied by the approximate counting structure will be $O(\varepsilon^{-1} n \log^{d+1} n)$. The query time will be

$$\begin{aligned} & O((Q_1(n) + \varepsilon^{-2} \log n \cdot Q_2(n)) \cdot \log(\varepsilon^{-1} \log n)) \\ & \equiv O((f(n) + \varepsilon^{-2} \log n \cdot f(n) \log n) \cdot \log(\varepsilon^{-1} \log n)) \\ & \equiv O((\log^{d-1} n + \varepsilon^{-2} \log^{d+1} n) \cdot \log(\varepsilon^{-1} \log n)) \\ & \equiv O(\varepsilon^{-2} \log^{d+1} n \cdot \log(\varepsilon^{-1} \log n)) \end{aligned}$$

APPENDIX for Section 4

9 Proof of Theorem 7

Primary Structure: Inspired by external memory tree structures with large fanout, our primary structure will be a segment tree [14] ST with fanout $f = \varepsilon^{-1} \log m$. Project the rectangles of R onto the x -axis. Let E_x be the set of endpoints of these projected intervals (each rectangle gets projected into an interval). Build an f -ary tree ST with the set E_x stored at the leaves. For each internal node $v \in ST$ we define its range on the x -axis, say $xrange(v)$, as the union of the $xrange(\cdot)$ of its children v_1, v_2, \dots, v_f , i.e., $xrange(v) = \bigcup_{i=1}^f xrange(v_i) = [x_l, x_r]$. For each internal node $v \in ST$ we also define $f + 1$ boundary slabs $b_1(v), b_2(v), \dots, b_{f+1}(v)$: $b_1(v) = x_l, b_{f+1}(v) = x_r$ and $\forall 1 < i < f + 1, b_i$ is the boundary separating $xrange(v_{i-1})$ and $xrange(v_i)$.

Consider a rectangle $r = [x_1, x_2] \times [y_1, y_2] \times (-\infty, z] \in R$. To store the rectangle r in ST , we start a tour from the root node of ST . Let v be the current node visited. There are two possibilities: (a) If $[x_1, x_2]$ does not intersect any of the boundary slabs, then we visit the child of v whose $xrange(\cdot)$ completely contains $[x_1, x_2]$. (b) Otherwise, the rectangle r is *assigned* to node v and broken into three disjoint rectangles as follows: Let x_1 lie in $xrange(v_i)$ and x_2 lie in $xrange(v_j)$. Then r is broken into a *left rectangle* $r_l = [x_1, b_{i+1}(v)) \times [y_1, y_2] \times (-\infty, z]$, *middle rectangle* $r_m =$

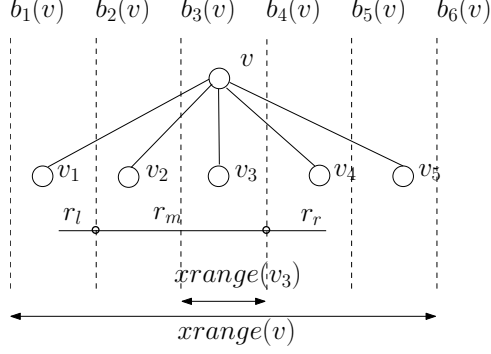


Figure 3: An internal node v in the segment tree ST .

$[b_{i+1}(v), b_j(v)] \times [y_1, y_2] \times (-\infty, z]$ and a *right rectangle* $r_r = (b_j(v), x_2] \times [y_1, y_2] \times (-\infty, z]$. The left (resp. right) rectangle is recursively stored in the subtree of v_i (resp. v_j). See Figure 3 for an illustration of a node $v \in ST$. The net effect is that each rectangle in R is assigned to at most two nodes per level in ST .

We augment each node $v \in ST$ with a couple of *secondary* data structures. Define R_v to be the set of rectangles assigned to node v . Next, define $R_{v_i} \subseteq R_v$ to be the set of rectangles completely crossing the $xrange(v_i)$ (In Figure 3 rectangle r completely crosses $xrange(v_2)$ and $xrange(v_3)$). For a given query point $q(q_x, q_y, q_z)$, if $q_x \in xrange(v_i)$, then the problem of approximately estimating the size of $R_v \cap q$ is reduced to the problem of approximately estimating the size of $R_{v_i} \cap q$. More importantly, the problem is now reduced to 3-sided rectangles in the yz -plane. Next, we present a secondary structure to efficiently estimate the size of $R_{v_i} \cap q$, under the assumption that $q_x \in xrange(v_i)$.

Secondary Structure: Project the rectangles R_{v_i} onto the y -axis. Let $E_y = (e_1, e_2, \dots, e_{2|R_{v_i}|})$ be the sorted sequence (in increasing order of y -coordinate) of the endpoints of these projected intervals. We divide the y -axis into *elementary segments* $(-\infty, e_1), [e_1, e_1], (e_1, e_2), [e_2, e_2], (e_2, e_3), \dots, (e_{2|R_{v_i}|-1}, e_{2|R_{v_i}|}), [e_{2|R_{v_i}|}, e_{2|R_{v_i}|}], (e_{2|R_{v_i}|}, \infty)$. For each elementary segment, say es , let $R_{v_i}^{es} \subseteq R_{v_i}$ be the set of rectangles completely crossing the segment es . The rectangles in $R_{v_i}^{es}$ are sorted in decreasing order of their *span* along the z -axis (rectangle $r_1 = [*] \times [*] \times (-\infty, z_1]$ has a larger span than rectangle $r_2 = [*] \times [*] \times (-\infty, z_2]$ if $z_1 > z_2$). For simplicity of notation, assume $R_{v_i}^{es}$ itself represents that sorted sequence. Storing the lists $R_{v_i}^{es}$ is not feasible as that would lead to a space consumption of $\Omega(m^2)$ for the entire data structure. Instead, we store an approximation $\mathcal{A}(R_{v_i}^{es})$, of the list $R_{v_i}^{es}$:

- The first $\lceil 1/\varepsilon \rceil$ entires in $\mathcal{A}(R_{v_i}^{es})$ will be the 1-st, 2-nd, \dots , $\lceil 1/\varepsilon \rceil$ -th largest span rectangle in $R_{v_i}^{es}$.
- The following entires in $\mathcal{A}(R_{v_i}^{es})$ will be the $\lceil \varepsilon^{-1}(1 + \varepsilon) \rceil$ -th, $\lceil \varepsilon^{-1}(1 + \varepsilon)^2 \rceil$ -th, $\lceil \varepsilon^{-1}(1 + \varepsilon)^3 \rceil$ -th, \dots largest span rectangle in $R_{v_i}^{es}$.

Query Algorithm: For a node $v \in ST$, if $q_x \in xrange(v_i)$ and $q_y \in es$, then we just need to find the predecessor of q_z in $\mathcal{A}(R_{v_i}^{es})$ to compute the approximate size of $R_v \cap q$. If the j -th element in $\mathcal{A}(R_{v_i}^{es})$ is the predecessor of q_z , then the rank of the j -th element in $R_{v_i}^{es}$ will be a valid approximation of $|R_v \cap q|$. If no predecessor is found, then 0 is a valid approximation of $|R_v \cap q|$. Finding the predecessor in $\mathcal{A}(R_{v_i}^{es})$ can be done in $O(\log(\varepsilon^{-1} \log m))$ time. For a given point $q(q_x, q_y, q_z)$, let Π be the path in ST from the root to the leaf node containing q_x . Note that

$\Pi = O(\log m / \log(\varepsilon^{-1} \log m))$. Roughly speaking, we have $O(\log(\varepsilon^{-1} \log m))$ time at each node in Π to search for the segment es containing q_y . Fortunately, this can be done by using the framework of fractional cascading [23]. The final value τ reported will be the sum of the approximate value returned at each node in Π .

Analysis: First we prove that $\tau \in [(1 - \varepsilon)k, k]$. Consider a node $v \in \Pi$ and let τ_v be the value returned by querying its secondary structure. Let the predecessor found in $\mathcal{A}(R_{v_i}^{es})$ be the i -th entry. We look at various ranges of i and handle each of them separately:

(i) When $i < \lceil 1/\varepsilon \rceil$: In this case we report the exact value of $|R_v \cap q|$.

(ii) When $i \geq \lceil 1/\varepsilon \rceil$: In this case we report $\tau_v = \lceil \varepsilon^{-1}(1 + \varepsilon)^{i - \lceil 1/\varepsilon \rceil} \rceil$. By our construction of $\mathcal{A}(R_{v_i}^{es})$, it should be clear that $\tau_v \leq |R_v \cap q|$. Now we show that $\tau \geq (1 - \varepsilon)|R_v \cap q|$.

$$\begin{aligned}
|R_v \cap q| &\leq \left\lceil \varepsilon^{-1}(1 + \varepsilon)^{i - \lceil 1/\varepsilon \rceil + 1} \right\rceil - 1 && \text{since the } i + 1\text{-th element in } \mathcal{A}(R_{v_i}^{es}) \text{ is the successor} \\
&\leq \left(\varepsilon^{-1}(1 + \varepsilon)^{i - \lceil 1/\varepsilon \rceil + 1} + 1 \right) - 1 \\
&= \varepsilon^{-1}(1 + \varepsilon)^{i - \lceil 1/\varepsilon \rceil + 1} \\
&\leq (1 + \varepsilon) \left\lceil \varepsilon^{-1}(1 + \varepsilon)^{i - \lceil 1/\varepsilon \rceil} \right\rceil \\
&= (1 + \varepsilon)\tau_v \\
&\text{which implies that } \tau_v \geq (1 - \varepsilon)|R_v \cap q|
\end{aligned}$$

Therefore, we have shown that $\tau_v \in [(1 - \varepsilon)|R_v \cap q|, |R_v \cap q|]$. Since, $\tau = \sum_{v \in \Pi} \tau_v$, we conclude that $\tau \in [(1 - \varepsilon)k, k]$.

Next, we analyze the size of our data structure. For a given R_{v_i} , the total size of all the approximate lists will be

$$\sum_{es} |\mathcal{A}(R_{v_i}^{es})| = O(\varepsilon^{-1}|R_{v_i}| \log m) = O(\varepsilon^{-1}|R_v| \log m)$$

For any given node $v \in ST$, since it has f children, the total size of all the approximate lists stored at node v will be:

$$O(f\varepsilon^{-1}|R_v| \log m) = O(\varepsilon^{-2}|R_v| \log^2 m)$$

The total size of all the secondary structures in ST will be:

$$\begin{aligned}
\sum_{v \in ST} O(\varepsilon^{-2}|R_v| \log^2 m) &\leq \varepsilon^{-2} \log^2 m \sum_{v \in ST} O(|R_v|) \\
&\leq (\varepsilon^{-2} \log^2 m) \cdot O(m \log m) && \text{since height of tree is bounded by } O(\log m) \\
&= O(\varepsilon^{-2} m \log^3 m)
\end{aligned}$$

Finally we analyze the query time. Since the height of the tree is $O\left(\frac{\log m}{\log(\varepsilon^{-1} \log m)}\right)$, finding the elementary segment es at all nodes in Π can be done in $O\left(\frac{\log m}{\log(\varepsilon^{-1} \log m)}\right) \times O(\log(\varepsilon^{-1} \log m)) = O(\log m)$ time. Then $O(\log(\varepsilon^{-1} \log m))$ time is spent to find the predecessor in $\mathcal{A}(R_{v_i}^{es})$ at each node in Π . Therefore, the overall query time is bounded by $O(\log m)$.

References

- [1] Peyman Afshani, Lars Arge, and Kasper Dalgaard Larsen. Orthogonal range reporting in three and higher dimensions. In *Proceedings of Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 149–158, 2009.
- [2] Peyman Afshani, Lars Arge, and Kasper Dalgaard Larsen. Orthogonal range reporting: query lower bounds, optimal structures in 3-d, and higher-dimensional improvements. In *Proceedings of Symposium on Computational Geometry (SoCG)*, pages 240–246, 2010.
- [3] Peyman Afshani, Lars Arge, and Kasper Green Larsen. Higher-dimensional orthogonal range reporting and rectangle stabbing in the pointer machine model. In *Proceedings of Symposium on Computational Geometry (SoCG)*, pages 323–332, 2012.
- [4] Peyman Afshani and Timothy M. Chan. On approximate range counting and depth. *Discrete & Computational Geometry*, 42(1):3–21, 2009.
- [5] Peyman Afshani and Timothy M. Chan. Optimal halfspace range reporting in three dimensions. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 180–186, 2009.
- [6] Peyman Afshani, Chris H. Hamilton, and Norbert Zeh. A general approach for cache-oblivious range reporting and approximate range counting. *Computational Geometry: Theory and Applications*, 43(8):700–712, 2010.
- [7] Pankaj K. Agarwal, Lars Arge, Jeff Erickson, Paolo Giulio Franciosa, and Jeffrey Scott Vitter. Efficient searching with linear constraints. *Journal of Computer and System Sciences (JCSS)*, 61(2):194–216, 2000.
- [8] Pankaj K. Agarwal, Lars Arge, Haim Kaplan, Eyal Molad, Robert Endre Tarjan, and Ke Yi. An optimal dynamic data structure for stabbing-semigroup queries. *SIAM Journal of Computing*, 41(1):104–127, 2012.
- [9] Pankaj K. Agarwal, Lars Arge, Jun Yang, and Ke Yi. I/O-efficient structures for orthogonal range-max and stabbing-max queries. In *Proceedings of European Symposium on Algorithms (ESA)*, pages 7–18, 2003.
- [10] Pankaj K. Agarwal, Lars Arge, and Ke Yi. An optimal dynamic interval stabbing-max data structure? In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 803–812, 2005.
- [11] Pankaj K. Agarwal and Jeff Erickson. Geometric range searching and its relatives. *Advances in Discrete and Computational Geometry*, pages 1–56, 1999.
- [12] Pankaj K. Agarwal, Sathish Govindarajan, and S. Muthukrishnan. Range searching in categorical data: Colored range searching on grid. In *Proceedings of European Symposium on Algorithms (ESA)*, pages 17–28, 2002.
- [13] Pankaj K. Agarwal and Jiri Matousek. Dynamic half-space range reporting and its applications. *Algorithmica*, 13(4):325–345, 1995.

- [14] Lars Arge and Jeffrey Scott Vitter. Optimal dynamic interval management in external memory. In *Proceedings of Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 560–569, 1996.
- [15] Boris Aronov and Sarel Har-Peled. On approximating the depth and related problems. *SIAM Journal of Computing*, 38(3):899–921, 2008.
- [16] Boris Aronov, Sarel Har-Peled, and Micha Sharir. On approximate halfspace range counting and relative epsilon-approximations. In *Proceedings of Symposium on Computational Geometry (SoCG)*, pages 327–336, 2007.
- [17] Timothy M. Chan. Random sampling, halfspace range reporting, and construction of ($\leq k$)-levels in three dimensions. *SIAM Journal of Computing*, 30(2):561–575, 2000.
- [18] Timothy M. Chan. Persistent predecessor search and orthogonal point location on the word RAM. *ACM Transactions on Algorithms*, 9(3):22:1–22:22, 2013.
- [19] Timothy M. Chan, Kasper Green Larsen, and Mihai Patrascu. Orthogonal range searching on the RAM, revisited. In *Proceedings of Symposium on Computational Geometry (SoCG)*, pages 1–10, 2011.
- [20] Timothy M. Chan and Bryan T. Wilkinson. Adaptive and approximate orthogonal range counting. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 241–251, 2013.
- [21] Bernard Chazelle. Filtering search: A new approach to query-answering. *SIAM Journal of Computing*, 15(3):703–724, 1986.
- [22] Bernard Chazelle. A functional approach to data structures and its use in multidimensional searching. *SIAM Journal of Computing*, 17(3):427–462, 1988.
- [23] Bernard Chazelle and Leonidas J. Guibas. Fractional cascading: I. A data structuring technique. *Algorithmica*, 1(2):133–162, 1986.
- [24] Prosenjit Gupta, Ravi Janardan, and Michiel H. M. Smid. Further results on generalized intersection searching problems: Counting, reporting, and dynamization. *Journal of Algorithms*, 19(2):282–317, 1995.
- [25] Prosenjit Gupta, Ravi Janardan, and Michiel H. M. Smid. Computational geometry: Generalized intersection searching. In *Handbook of Data Structures and Applications*. 2004.
- [26] Sarel Har-Peled and Micha Sharir. Relative (p, ϵ) -approximations in geometry. *Discrete & Computational Geometry*, 45(3):462–496, 2011.
- [27] Joseph JáJá, Christian Worm Mortensen, and Qingmin Shi. Space-efficient and fast algorithms for multidimensional dominance reporting and counting. In *Algorithms and Computation, 15th International Symposium, ISAAC 2004, Hong Kong, China, December 20-22, 2004, Proceedings*, pages 558–568, 2004.
- [28] Ravi Janardan and Mario A. Lopez. Generalized intersection searching problems. *International Journal of Computational Geometry and Applications*, 3(1):39–69, 1993.

- [29] Haim Kaplan, Eyal Molad, and Robert Endre Tarjan. Dynamic rectangular intersection with priorities. In *Proceedings of ACM Symposium on Theory of Computing (STOC)*, pages 639–648, 2003.
- [30] Haim Kaplan, Edgar Ramos, and Micha Sharir. Range minima queries with respect to a random permutation, and approximate range counting. *Discrete & Computational Geometry*, 45(1):3–33, 2011.
- [31] Haim Kaplan, Natan Rubin, Micha Sharir, and Elad Verbin. Counting colors in boxes. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 785–794, 2007.
- [32] Haim Kaplan and Micha Sharir. Randomized incremental constructions of three-dimensional convex hulls and planar voronoi diagrams, and approximate range counting. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 484–493, 2006.
- [33] Kasper Green Larsen and Rasmus Pagh. I/O-efficient data structures for colored range and prefix reporting. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 583–592, 2012.
- [34] Kasper Green Larsen and Freek van Walderveen. Near-optimal range reporting structures for categorical data. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 265–276, 2013.
- [35] Jirí Matousek. Reporting points in halfspaces. *Computational Geometry: Theory and Applications*, 2:169–186, 1992.
- [36] Jirí Matousek. Range searching with efficient hierarchical cutting. *Discrete & Computational Geometry*, 10:157–182, 1993.
- [37] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.
- [38] Yakov Nekrich. Efficient range searching for categorical and plain data. *ACM Transactions on Database Systems (TODS)*, 39(1):9, 2014.
- [39] Mihai Patrascu. Lower bounds for 2-dimensional range counting. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 11-13, 2007*, pages 40–46, 2007.
- [40] Mihai Patrascu. Unifying the landscape of cell-probe lower bounds. *SIAM Journal of Computing*, 40(3):827–847, 2011.
- [41] Octavian Procopiuc, Pankaj K. Agarwal, Lars Arge, and Jeffrey Scott Vitter. Bkd-tree: A dynamic scalable kd-tree. In *Proceedings of Symposium on Advances in Spatial and Temporal Databases (SSTD)*, pages 46–65, 2003.
- [42] Qingmin Shi and Joseph JáJá. Optimal and near-optimal algorithms for generalized intersection reporting on pointer machines. *Information Processing Letters (IPL)*, 95(3):382–388, 2005.